



US009355167B2

(12) **United States Patent**
Gomes et al.

(10) **Patent No.:** **US 9,355,167 B2**
(45) **Date of Patent:** **May 31, 2016**

(54) **SYSTEMS AND METHODS FOR THE
DISTRIBUTED CATEGORIZATION OF
SOURCE DATA**

(71) Applicant: **California Institute of Technology,**
Pasadena, CA (US)

(72) Inventors: **Ryan Gomes**, Santa Monica, CA (US);
Peter Welinder, San Diego, CA (US);
Andreas Krause, Zurich (CH); **Pietro**
Perona, Altadena, CA (US)

(73) Assignee: **California Institute of Technology,**
Pasadena, CA (US)

(*) Notice: Subject to any disclaimer, the term of this
patent is extended or adjusted under 35
U.S.C. 154(b) by 149 days.

(21) Appl. No.: **13/897,229**

(22) Filed: **May 17, 2013**

(65) **Prior Publication Data**

US 2014/0289246 A1 Sep. 25, 2014

Related U.S. Application Data

(60) Provisional application No. 61/648,965, filed on May
18, 2012, provisional application No. 61/663,138,
filed on Jun. 22, 2012.

(51) **Int. Cl.**
G06F 17/30 (2006.01)

(52) **U.S. Cl.**
CPC **G06F 17/30598** (2013.01); **G06F 17/30525**
(2013.01)

(58) **Field of Classification Search**
USPC 707/705, 778
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

6,636,843 B2 * 10/2003 Doddi G02B 5/18
706/12
6,897,875 B2 * 5/2005 Zhang G06K 9/6253
345/586

7,610,130 B1 10/2009 Dixon et al.
7,809,722 B2 10/2010 Gokturk et al.
7,987,186 B1 7/2011 Joshi
8,041,568 B2 * 10/2011 Strobe G06Q 10/103
704/257

8,418,249 B1 4/2013 Nucci et al.
2012/0158620 A1 6/2012 Paquet et al.
2012/0221508 A1 8/2012 Chaturvedi et al.

(Continued)

OTHER PUBLICATIONS

Antoniak, "Mixtures of Dirchlet Processes with Applications to
Bayesian Nonparametric Problems", The Annals of Statistics, 1974,
vol. 2, No. 6, pp. 1152-1174.

(Continued)

Primary Examiner — Kim Nguyen

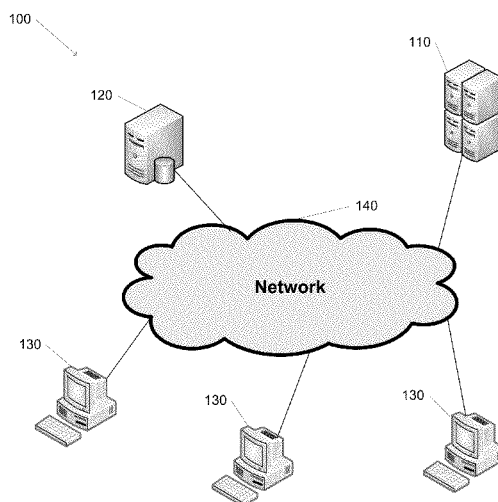
(74) *Attorney, Agent, or Firm* — KPPB LLP

(57)

ABSTRACT

Systems and methods for the crowdsourced clustering of data
items in accordance embodiments of the invention are dis-
closed. In one embodiment of the invention, a method for
determining categories for a set of source data includes
obtaining a set of source data, determining a plurality of
subsets of the source data, where a subset of the source data
includes a plurality of pieces of source data in the set of source
data, generating a set of pairwise annotations for the pieces of
source data in each subset of source data, clustering the set of
source data into related subsets of source data based on the
sets of pairwise labels for each subset of source data, and
identifying a category for each related subset of source data
based on the clusterings of source data and the source data
metadata for the pieces of source data in the group of source
data.

22 Claims, 5 Drawing Sheets



(56)

References Cited

U.S. PATENT DOCUMENTS

2013/0024457 A1 1/2013 Chua et al.
 2013/0080422 A1* 3/2013 Pan G06Q 30/00
 707/722
 2013/0097164 A1 4/2013 Welinder et al.

OTHER PUBLICATIONS

- Attias, "A Variational Bayesian Framework for Graphical Models", NIPS, 1999, pp. 209-215.
- Bennett, "Using Asymmetric Distributions to Improve Classifier Probabilities: A Comparison of New and Standard Parametric Methods", Technical report, Carnegie Mellon University, 2002, 24 pgs.
- Berg et al., "Automatic Attribute Discovery and Characterization from Noisy Web Data", Computer Vision—ECCV 2010, pp. 663-676.
- Beygelzimer et al., "Importance Weighted Active Learning", Proceedings of the 26th International Conference on Machine Learning, 2009, 8 pgs.
- Bourdev et al., "Poselets: Boyd Part Detectors Trained Using 3D Human Pose Annotations", ICCV, 2009, 42 pgs.
- Byrd et al., "A Limited Memory Algorithm for Bound Constrained Optimization", SIAM Journal on Scientific and Statistical Computing, 1995, vol. 16, No. 5, pp. 1190-1208.
- Cohn et al., "Active Learning with Statistical Models", Journal of Artificial Intelligence Research, 1996, vol. 4, pp. 129-145.
- Dalal et al., "Histograms of Oriented Gradients for Human Detection", ICCV, 2005, 8 pgs.
- Dankert et al., "Automated Monitoring and Analysis of Social Behavior in Drosophila", Nat. Methods, Apr. 2008, vol. 6, No. 4, 17 pgs.
- Dasgupta et al., "Hierarchical Sampling for Active Learning", ICML, 2008, 8 pgs.
- Dawid et al., "Maximum Likelihood Estimation of Observer Error-rates using the EM Algorithm", J. Roy. Statistical Society, Series C, 1979, vol. 28, No. 1, pp. 20-28.
- Deng et al., "ImageNet: A Large-Scale Hierarchical Image Database", CVPR, 2009, 8 pgs.
- Dollar et al., "Cascaded Pose Regression", CVPR, 2010, pp. 1078-1085.
- Dollar et al., "Pedestrian Detection: An Evaluation of the State of the Art", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 20 pgs.
- Erkanli et al., "Bayesian semi-parametric ROC analysis", Statistics in Medicine, 2006, vol. 25, pp. 3905-3928.
- Fei-Fei et al., "A Bayesian Hierarchical Model for Learning Natural Scene Categories", CVPR, IEEE Computer Society, 2005, pp. 524-531.
- Frank et al., "UCI Machine Learning Repository", 2010, 3 pgs.
- Fuchs et al., "Randomized tree ensembles for object detection in computational pathology", Lecture Notes in Computer Science, ISVC, 2009, vol. 5875, pp. 367-378.
- Gionis et al., "Clustering Aggregation", ACM Transactions on Knowledge Discovery from Data, 2007, vol. 1, 30 pgs.
- Gomes, et al., "Crowdclustering", Technical Report, Caltech 20110628-202526159, Jun. 2011, 14 pgs.
- Gomes et al., "Crowdclustering", Technical Report, Caltech, 2011, pp. 558-561.
- Gu et al., "Bayesian bootstrap estimation of ROC curve", Statistics in Medicine, 2008, vol. 27, pp. 5407-5420.
- Hellmich et al., "Bayesian Approaches to Meta-analysis of ROC Curves", Med. Decis. Making, Jul.-Sep. 1999, vol. 19, pp. 252-264.
- Jaakkola et al., "A variational approach to Bayesian logistic regression models and their extensions", Source unknown, Aug. 13, 1996, 10 pgs.
- Kruskal, "Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis", Psychometrika, Mar. 1964, vol. 29, No. 1, pp. 1-27.
- Kurihara et al., "Accelerated Variational Dirichlet Process Mixtures", Advances in Neural Information Processing Systems, 2007, 8 pgs.
- Li et al., "Solving consensus and Semi-supervised Clustering Problems Using Nonnegative Matrix Factorization", ICDM, IEEE computer society 2007, pp. 577-582.
- Little et al., "Exploring Iterative and Parallel Human Computational Processes", HCOMP, 2010, pp. 68-76.
- Little et al., "TurKit: Tools for Iterative Tasks on Mechanical Turk", HCOMP, 2009, pp. 29-30.
- Maceachern et al., "Estimating Mixture of Dirichlet Process Models", Journal of Computational and Graphical Statistics, Jun. 1998, vol. 7, No. 2, pp. 223-238.
- Mackay, "Information-Based objective Functions for Active Data Selection", Neural Computation, 1992, vol. 4, pp. 590-604.
- Martinez-Munoz et al., "Dictionary-Free Categorization of Very Similar Objects via Stacked Evidence Trees", Source unknown, 2009, 8 pgs.
- Meila, "Comparing Clusterings by the Variation of Information", Learning theory and Kernel machines: 16th Annual Conference of Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, 31 pgs.
- Monti et al., "Consensus Clustering: A Resampling-Based Method for Class Discovery and Visulation of Gene Expression Microarray Data", Machine Learning, 2003, vol. 52, pp. 91-118.
- Neal, "MCMC Using Hamiltonian Dynamics", Handbook of Markov Chain Monte Carlo, 2010, pp. 113-162.
- Nigam et al., "Text Classification from labeled and Unlabeled Documents Using EM", Machine Learning, 2000, vol. 39, No. 2/3, pp. 103-134.
- Platt, "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods", Advances in Large Margin Classifiers, 1999, MIT Press, pp. 61-74.
- Raykar et al., "Supervised Learning from Multiple Experts: Whom to trust when everyone lies a bit", ICML, 2009, 8 pgs.
- Russell et al., "LabelMe: A Database and Web-Based Tool for Image Annotation", Int. J. Computer. Vis., 2008, vol. 77, pp. 157-173.
- Seeger, "Learning with labeled and unlabeled data", Technical Report, University of Edinburgh, 2002, 62 pgs.
- Sheng et al., "Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers", KDD, 2008, 9 pgs.
- Smyth et al., "Inferring Ground Truth from Subjective Labeling of Venus Images", NIPS, 1995, 8 pgs.
- Snow et al., "Cheap and Fast—But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks", EMNLP, 2008, 10 pgs.
- Sorokin et al., "Utility data annotation with Amazon Mechanical Turk", First IEEE Workshop on Internet Vision at CVPR '08, 2008, 8 pgs.
- Spain et al., "Some objects are more equal than others: measuring and predicting importance", ECCV, 2008, 14 pgs.
- Strehl et al., "Cluster Ensembles—A Knowledge Reuse Framework for Combining Multiple Partitions", Journal of Machine Learning Research, 2002, vol. 3, pp. 583-617.
- Tong et al., "Support Vector Machine Active Learning with Applications to Text Classification", Journal of Machine Learning Research, 2001, pp. 45-66.
- Vijayanarasimhan et al., "Large-Scale Live Active Learning: Training Object Detectors with Crawled Data and Crowds", CVPR, 2001, pp. 1449-1456.
- Vijayanarasimhan et al., "What's It Going to Cost You?: Predicting Effort vs. Informativeness for Multi-Label Image Annotations", CVPR, 2009, pp. 2262-2269.
- Von Ahn et al., "Labeling Images with a Computer Game", SIGCHI conference on Human factors in computing systems, 2004, pp. 319-326.
- Von Ahn et al., "reCAPTCHA: Human-Based Character Recognition via Web Security Measures", Science, 2008, vol. 321, No. 5895, pp. 1465-1468.
- Vondrick et al., "Efficiently Scaling up Video Annotation with Crowdsourced Marketplaces", ECCV, 2010, pp. 610-623.
- Welinder et al., "Caltech-UCSD Birds 200", Technical Report CNS-TR-2010-001, 2001, 15 pgs.
- Welinder et al., "Online crowdsourcing: rating annotators and obtaining cost-effective labels", IEEE Conference on Computer Vision and Pattern Recognition Workshops (ACVHL), 2010, pp. 25-32.

(56)

References Cited

OTHER PUBLICATIONS

Welinder et al., "The Multidimensional Wisdom of Crowds", Neural Information Processing Systems Conference (HIPS), 2010, pp. 1-9.
Whitehill et al., "Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise", NIPS, 2009, 9 pgs.
Zhu, "Semi-Supervised Learning Literature Survey", Technical report, University of Wisconsin-Madison, 2008, 60 pgs.

Bitton, E. "Geometric Models for Collaborative Search and Filtering", 2011, 186 pages.
Ertekin, S. et al., "Learning to predict the wisdom of crowds", arXiv preprint arXiv:1204.3611 v1, Apr. 16, 2012, 8 Pages.
Ertekin, S. et al., "Wisely Using a Budget for Crowdsourcing", OR 392-12. Massachusetts Institute of Technology, Apr. 2012, 31 pages.
Zhao, L. et al., "Robust Active Learning Using Crowdsourced Annotations for Activity Recognition", In Human Computation: Papers from the 2011 All Workshop., Jan. 2011, pp. 74-79.

* cited by examiner

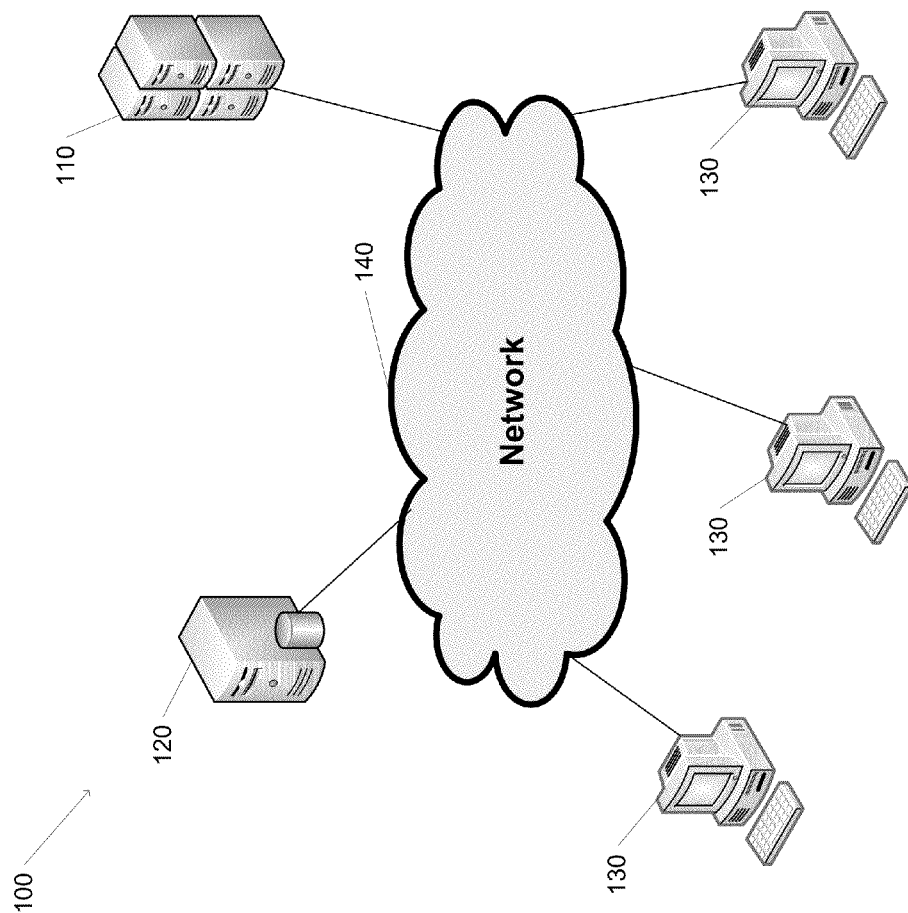


FIG. 1

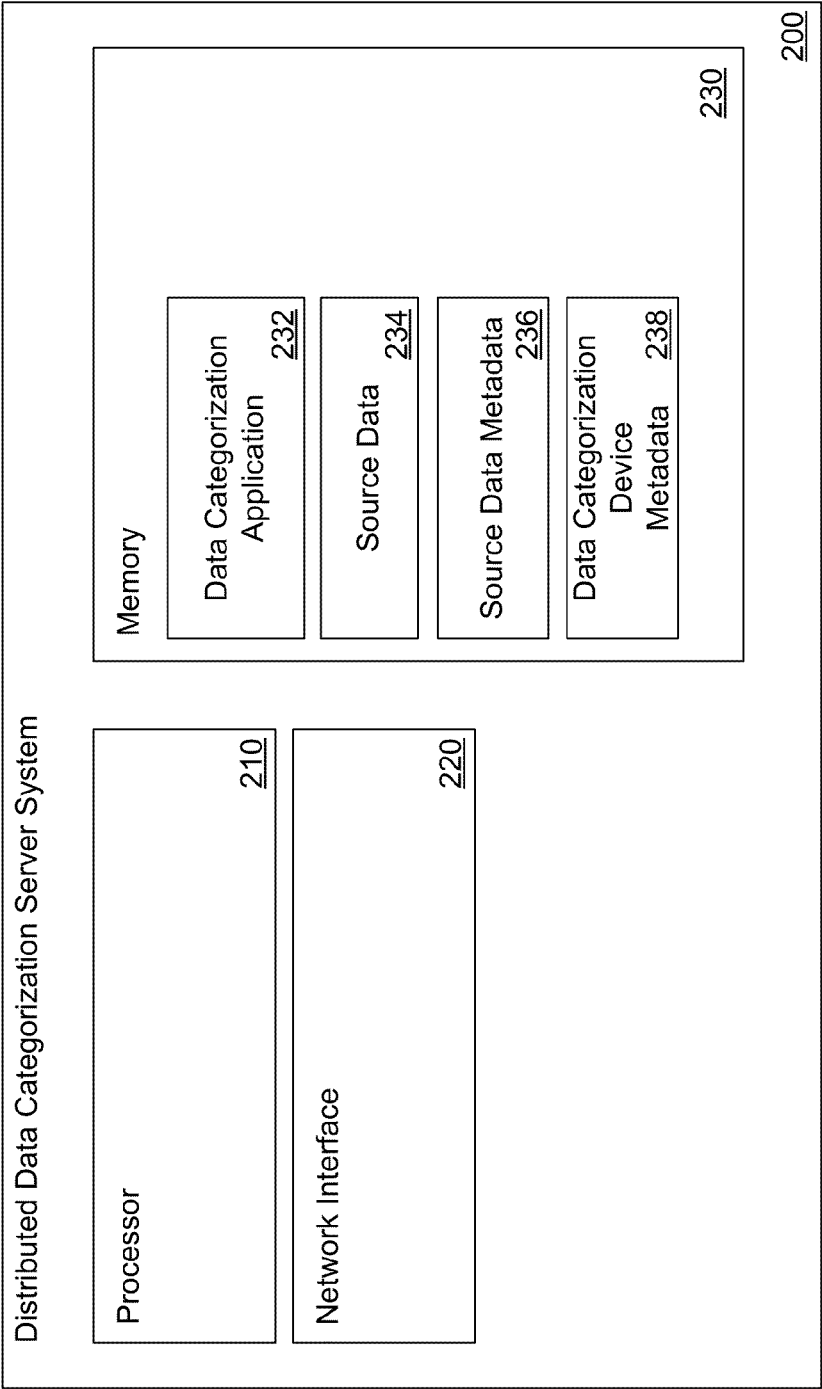
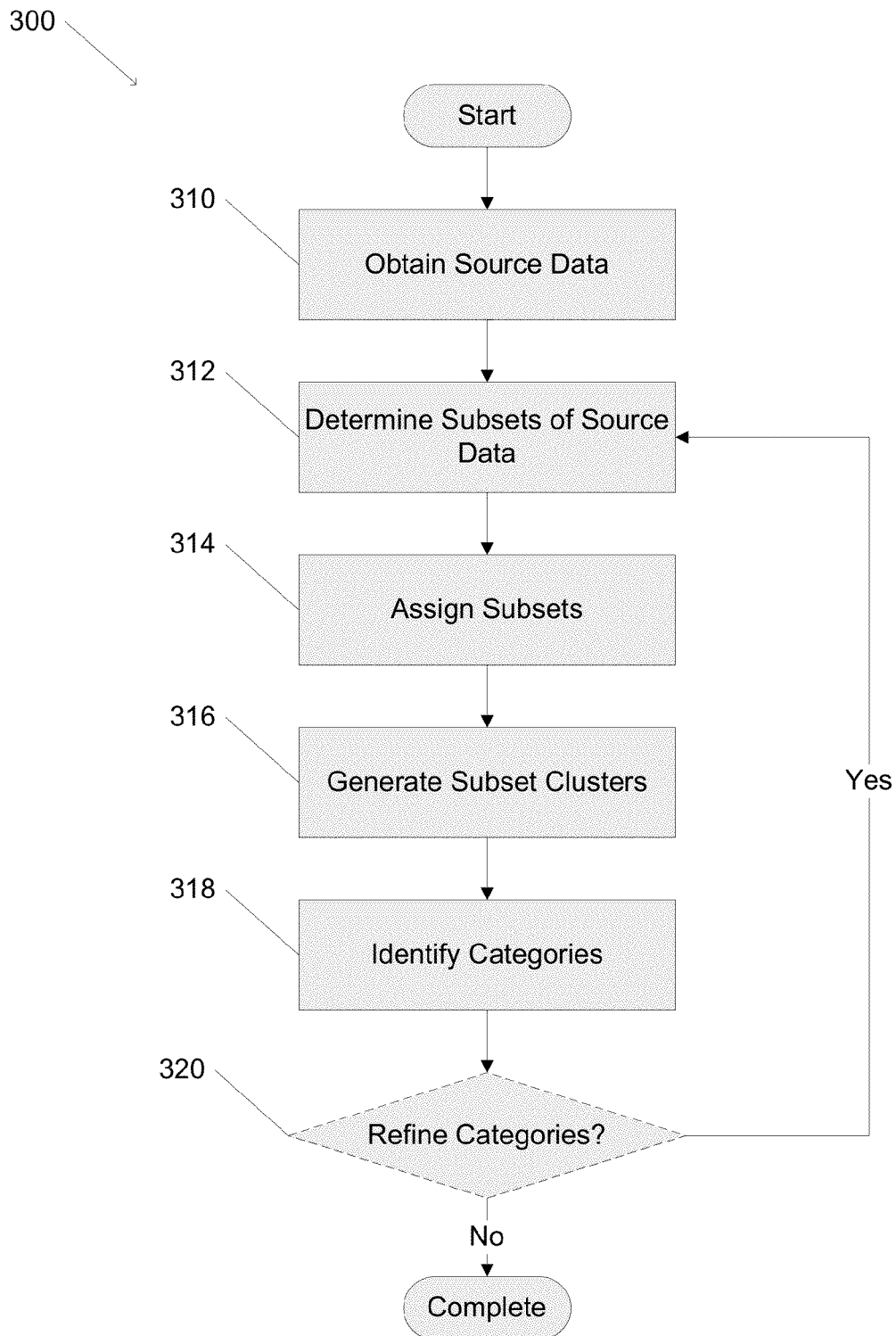
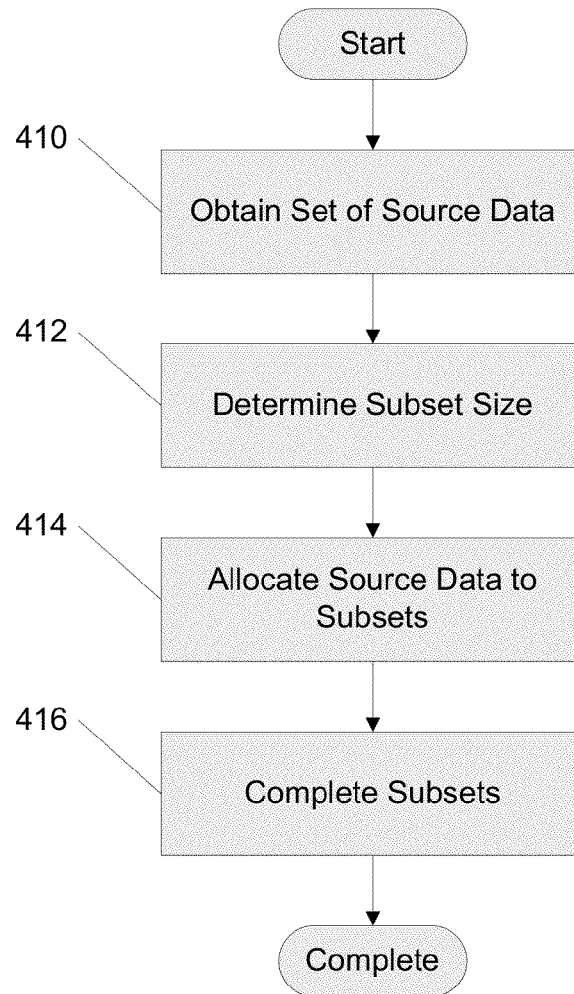
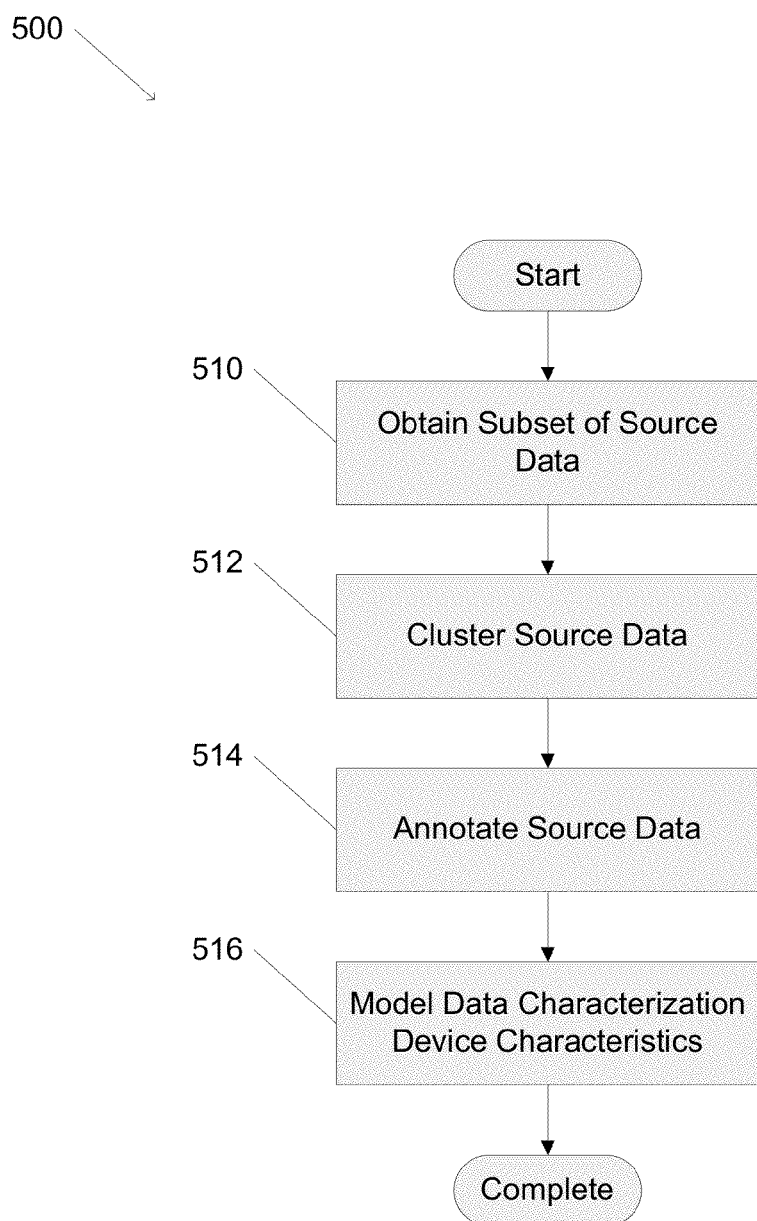


FIG. 2

**FIG. 3**

400

**FIG. 4**

**FIG. 5**

1

SYSTEMS AND METHODS FOR THE DISTRIBUTED CATEGORIZATION OF SOURCE DATA

CROSS-REFERENCE TO RELATED APPLICATIONS

The current application claims priority to U.S. Provisional Patent Application No. 61/648,965, titled "Crowdclustering" to Gomes et al. and filed May 18, 2012 and U.S. Provisional Patent Application No. 61/663,138, titled "Method for Combining Human and Machine Computation for Classification and Regression Tasks" to Welinder et al. and filed Jun. 22, 2012. The disclosures of U.S. Provisional Patent Application Nos. 61/648,965 and 61/663,138 are hereby incorporated by reference in their entirety.

STATEMENT OF FEDERAL FUNDING

This invention was made with government support under Grant No. IIS0413312 awarded by the National Science Foundation and under Grant No. N00014-06-1-0734 and Grant No. N00014-10-1-0933 awarded by the Office of Naval Research. The government has certain rights in the invention.

FIELD OF THE INVENTION

The present invention is generally related to data c and more specifically the distributed categorization of sets of data.

BACKGROUND OF THE INVENTION

Amazon Mechanical Turk is a service provided by Amazon.com of Seattle, Wash. Amazon Mechanical Turk provides the ability to submit tasks and have a human complete the task in exchange for a monetary reward for completing the task.

SUMMARY OF THE INVENTION

Systems and methods for the crowdsourced clustering of data items in accordance embodiments of the invention are disclosed. In one embodiment of the invention, a method for determining categories for a set of source data includes obtaining a set of source data using a distributed data categorization server system, where a piece of source data includes source data metadata describing attributes of the piece of source data, determining a plurality of subsets of the source data using the distributed data categorization server system, where a subset of the source data includes a plurality of pieces of source data in the set of source data, generating a set of pairwise annotations for the pieces of source data in each subset of source data using the distributed data categorization server system, where a pairwise annotation indicates when a first piece of source data in a pair of pieces of source data in the subset of source data is similar to a second piece of source data in the pair of pieces of source data, clustering the set of source data into related subsets of source data based on the sets of pairwise labels for each subset of source data using the distributed data categorization server system, and identifying a category for each related subset of source data based on the clusterings of source data and the source data metadata for the pieces of source data in the group of source data using the distributed data categorization server system.

In another embodiment of the invention, determining categories for a set of source data further includes generating a taxonomy based on the identified categories and the set of

2

source data using the distributed data categorization server system, where the taxonomy includes relationships between the identified categories and the pieces of source data in the set of source data.

5 In an additional embodiment of the invention, a category in the taxonomy includes one or more attributes of the pieces of source data associated with the category in the taxonomy.

In yet another additional embodiment of the invention, determining categories for a set of source data further includes iteratively identifying sub-categories for at least one identified category based on the pieces of source data associated with the identified category using the distributed data categorization server system.

10 In still another additional embodiment of the invention, the at least one identified category is selected based on the attributes of the pieces of source data associated with the identified category and the identified sub-categories include at least one attribute from a piece of source data associated with the sub-category that is not present in the identified category.

20 In yet still another additional embodiment of the invention, determining categories for a set of source data further includes generating instruction data using the distributed data categorization server system, where the instruction data describes the attributes of the pieces of the source data that should be used in generating the set of pairwise annotations.

In yet another embodiment of the invention, the instruction data is generated based on the attributes of the pieces of source data in the set of source data.

30 In still another embodiment of the invention, generating a set of pairwise annotations for the pieces of source data in each subset of source data using the distributed data categorization server system is based on data characterization device metadata, where the data characterization device metadata describes anticipated annotations based on the pieces of source data.

In yet still another embodiment of the invention, clustering the set of source data into related subsets of source data further includes generating a model including a set of points representing the pieces of source data in a Euclidian space using the distributed data categorization server system and clustering the set of points within the Euclidian space based on the set of pairwise annotations using the distributed data categorization server system.

45 In yet another additional embodiment of the invention, determining categories for a set of source data further includes estimating the number of clusters within the Euclidian space using the distributed data categorization server system.

50 In still another additional embodiment of the invention, determining a plurality of subsets further includes determining a subset size using the distributed data categorization server system, where the subset size is a measure of the number of pieces of source data assigned to a subset and deterministically allocating the pieces of source data to the determined subsets using the distributed data categorization server system.

55 In yet still another additional embodiment of the invention, determining categories for a set of source data further includes allocating additional pieces of source data to the subsets using the distributed data categorization server system, where the additional pieces of source data are sampled without replacement from the set of source data not already assigned to the subset.

60 Still another embodiment of the invention includes a distributed data categorization server system including a processor and a memory configured to store a data categorization

application, wherein the data categorization application configures the processor to obtain a set of source data, where a piece of source data includes source data metadata describing attributes of the piece of source data, determine a plurality of subsets of the source data, where a subset of the source data includes a plurality of pieces of source data in the set of source data, generate a set of pairwise annotations for the pieces of source data in each subset of source data, where a pairwise annotation indicates when a first piece of source data in a pair of pieces of source data in the subset of source data is similar to a second piece of source data in the pair of pieces of source data, cluster the set of source data into related subsets of source data based on the sets of pairwise labels for each subset of source data, and identify a category for each related subset of source data based on the clusterings of source data and the source data metadata for the pieces of source data in the group of source data.

In yet another additional embodiment of the invention, the data categorization application further configures the processor to generate a taxonomy based on the identified categories and the set of source data, where the taxonomy includes relationships between the identified categories and the pieces of source data in the set of source data.

In still another additional embodiment of the invention, a category in the taxonomy includes one or more attributes of the pieces of source data associated with the category in the taxonomy.

In yet still another additional embodiment of the invention, the data categorization application further configures the processor to iteratively identify sub-categories for at least one identified category based on the pieces of source data associated with the identified category.

In yet another embodiment of the invention, the at least one identified category is selected based on the attributes of the pieces of source data associated with the identified category and the identified sub-categories include at least one attribute from a piece of source data associated with the sub-category that is not present in the identified category.

In still another embodiment of the invention, the data categorization application further configures the processor to generate instruction data, where the instruction data describes the attributes of the pieces of the source data that should be used in generating the set of pairwise annotations.

In yet still another embodiment of the invention, the instruction data is generated based on the attributes of the pieces of source data in the set of source data.

In yet another additional embodiment of the invention, generating a set of pairwise annotations for the pieces of source data in each subset of source data using the distributed data categorization server system is based on data characterization device metadata, where the data characterization device metadata describes anticipated annotations based on the pieces of source data.

In still another additional embodiment of the invention, the data categorization application further configures the processor to cluster the set of source data into related subsets of source data by generating a model including a set of points representing the pieces of source data in a Euclidian space and clustering the set of points within the Euclidian space based on the set of pairwise annotations.

In yet still another additional embodiment of the invention, the data categorization application further configures the processor to estimate the number of clusters within the Euclidian space.

In yet another additional embodiment of the invention, the data categorization application further configures the process to determine a plurality of subsets by determining a subset

size, where the subset size is a measure of the number of pieces of source data assigned to a subset and deterministically allocating the pieces of source data to the determined subsets.

In still another additional embodiment of the invention, the data categorization application further configures the processor to allocate additional pieces of source data to the subsets, where the additional pieces of source data are sampled without replacement from the set of source data not already assigned to the subset.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 conceptually illustrates a distributed data categorization system in accordance with an embodiment of the invention.

FIG. 2 conceptually illustrates a data categorization device in accordance with an embodiment of the invention.

FIG. 3 is a flow chart conceptually illustrating a process for the categorization of source data in accordance with an embodiment of the invention.

FIG. 4 is a flow chart conceptually illustrating a process for determining subsets of source data in accordance with an embodiment of the invention.

FIG. 5 is a flow chart conceptually illustrating a process for modeling worker behavior in the categorization of source data in accordance with an embodiment of the invention.

DETAILED DESCRIPTION

Turning now to the drawings, systems and methods for distributed categorization of source data in accordance with embodiments of the invention are illustrated. In a variety of applications including, but not limited to, medical diagnosis, surveillance verification, performing data de-duplication, transcribing audio recordings, or researching data details, a large variety of source data, such as image data, audio data, and text data, can be generated and/or obtained. By categorizing these pieces of source data, particular portions of the source data can be identified for particular purposes and/or additional analysis. Systems and methods for annotating source data that can be utilized in accordance with embodiments of the invention are disclosed in U.S. patent application Ser. No. 13/651,108, titled "Systems and Methods for Distributed Data Annotation" to Welinder et al. and filed Oct. 12, 2012, the entirety of which is hereby incorporated by reference. However, particularly in applications having a large number of source data, determining what categories are present in the source data can be difficult. It may not be realistic to expect a single person or machine to look at all images and determine categories for the set of source data, as it may not be time and/or cost effective to have a single source of categorizations. Likewise, a single person or machine may not be able to identify every category present within the set of source data. Additionally, individual sources of category annotations, whether untrained sources or expert sources, might not agree on the criteria used to define categories and may not even agree on the number of categories that are present within the set of source data.

Distributed data categorization server systems in accordance with embodiments of the invention are configured to determine subsets of a set of source data and distribute the subsets of source data to a variety of data categorization devices that are configured to identify clusters containing similar source data within the subsets and annotate the pieces of source data with the identified cluster information. Data categorization devices include human annotators, machine-

5

based categorization devices, and/or a combination of machine and human categorization as appropriate to the requirements of specific applications in accordance with embodiments of the invention. The data categorization devices are configured to cluster subsets of source data based on a variety of categorization criteria, where the source data in a cluster belong to the same category according to the categorization criteria. In several embodiments, the clustering of data is a human intelligence task where human annotators are asked to pick pairs of source data that they consider to be most similar from a set of examples of source data. As can readily be appreciated, annotators can utilize any of a variety of characteristics of the source data to define similarly. For example, one annotator may cluster images of objects based upon color and another annotator may cluster the same images based on shape. The different characteristics utilized during annotation are typically unknown to the distributed data categorization and so analysis of the clusters reveals different attributes of the source data and semantic information concerning the manner in which the source data can be categorized. Furthermore, aggregation of information across a number of annotators can capture information concerning different useful ways in which users may categorize source data. In this way, source data can be annotated using in the ways that are most meaningful to users instead of being restricted to annotation in accordance with a predefined taxonomy. In a variety of embodiments, instruction data is provided along with the subsets of source data indicating the attributes of the pieces of source data that should be analyzed in the clustering of the pieces of source data. The distributed data categorization server system is configured to receive the different clusters of source data from a variety of data categorization devices and identify categories within the set of source data based on the annotations associated with the source data. The categories can be defined using the attributes of the pieces of source data as appropriate to the requirements of specific applications in accordance with embodiments of the invention. In a variety of embodiments, the distributed data categorization server system is configured to determine metadata describing the characteristics of the data categorization devices based on the received annotated pieces of source data. This data categorization device metadata describes the annotations received from the data categorization device and can be utilized to anticipate annotations provided by the data categorization device, in the determination of which data categorization devices to distribute subsets of source data, and in the calculation of rewards (such as monetary compensation) to allocate to particular data categorization devices.

In many embodiments, the distributed data categorization server system is configured to identify categories that may contain additional sub-categories and distribute those source data associated with an identified category for additional clustering by the data categorization devices. In this way, broad categories can be identified for a set of source data and additional categories can be iteratively identified within the source data. In a number of embodiments, the iterative categorization of source data is utilized to construct a taxonomy (or any other structure representing entities and the relationships between the entities) where the categories and sub-categories describe the relationships between the pieces of source data in the set of source data. As additional sub-categories are iteratively identified within subsets of the set of source data, additional relationships between the identified categories, sub-categories, and pieces of source data are incorporated into the taxonomy. Although specific taxonomy-based approaches for expressing the relationships

6

between categories and source data with increased specificity are discussed above, any of a variety of techniques can be utilized to categorize source data including techniques that involve a single pass or multiple passes by the same set of data categorization devices or different sets of data categorization devices as appropriate to the requirements of specific applications in accordance with embodiments of the invention.

Although the above is described with respect to distributed data categorization server systems and data categorization devices, the data categorization devices can be implemented using the distributed data categorization server system as appropriate to the requirements of specific applications in accordance with embodiments of the invention. Systems and methods for distributed categorization of source data in accordance with embodiments of the invention are discussed further below.

Distributed Data Categorization Systems

Distributed data categorization systems in accordance with embodiments of the invention are configured to distribute subsets of a set of source data to a variety of data categorization devices and, based on the results obtained from the data categorization devices, identify categories of source data within the set of source data. A conceptual illustration of a distributed data categorization system in accordance with an embodiment of the invention is shown in FIG. 1. Distributed data categorization system **100** includes distributed data categorization server system **110** connected to source data database **120** and one or more data categorization devices **130** via network **140**. In many embodiments, distributed data categorization server system **110** and/or source data database **120** are implemented using a single server. In a variety of embodiments, distributed data categorization server system **110** and/or source data database **120** are implemented using a plurality of servers. In many embodiments, data categorization devices **130** are implemented utilizing distributed data categorization server system **110** and/or source data database **120**. Network **140** can be one or more of a variety of networks, including, but not limited to, wide-area networks, local area networks, and/or the Internet as appropriate to the requirements of specific applications in accordance with embodiments of the invention.

Distributed data categorization system **110** is configured to obtain pieces of source data and store the pieces of source data using source data database **120**. Source data database **120** can obtain source data from any of a variety of sources, including content sources, customers, and any of a variety of providers of source data as appropriate to the requirements of specific applications in accordance with embodiments of the invention. In a variety of embodiments, source data database **120** includes one or more references (such as a uniform resource locator) to source data that is stored in a distributed fashion. Source data database **120** includes one or more sets of source data to be categorized using distributed data categorization server system **110**. A set of source data includes one or more pieces of source data including, but not limited to, image data, audio data, signal data, and text data. In several embodiments, one or more pieces of source data in source data database **120** includes source data metadata describing attributes of the piece of source data. Distributed data categorization server system **110** is further configured to generate subsets of source data and distribute the subsets of source data to one or more data categorization devices **130**. Data categorization devices **130** transmit annotated source data to distributed data categorization server system **110**. Based on the annotated source data, distributed data categorization server system **110** is configured to identify categories describing the pieces of source data. In many embodiments, distributed data categorization

rization server system **110** is configured to determine the characteristics of the data categorization devices **130** based on the received annotations. The characteristics of data categorization devices **130** can be utilized by distributed data categorization server system **110** to determine which data categorization devices **130** will receive pieces of source data and/or determine rewards (or other compensation) for annotating pieces of source data. In a number of embodiments, distributed data categorization server system **110** is configured to identify categories of source data that may contain additional sub-categories within the source data and (iteratively) distribute the identified categories of source data to data categorization devices **130** to identify additional sub-categories within the particular set of source data.

Data categorization devices **130** are configured to cluster pieces of source data according to categorization criteria and annotate the pieces of source data based on the clustering via metadata associated with the pieces of source data. Data categorization devices **130** include, but are not limited to, human annotators, machine annotators, and emulations of human annotators performed using machines. Human annotators can constitute any human-generated annotators, including users performing human intelligence tasks via a service such as the Amazon Mechanical Turk service provided by Amazon.com, Inc. In the illustrated embodiment, data categorization devices **130** are illustrated as personal computers configured using appropriate software. In various embodiments, data categorization devices can include (but are not limited to) tablet computers, mobile phone handsets, software running on distributed data categorization server system **110**, and/or any of a variety of network-connected devices as appropriate to the requirements of specific applications in accordance with embodiments of the invention. In several embodiments, data categorization devices **130** provide a user interface and an input device configured to allow a user to view the pieces of source data received by the data categorization device and provide annotations (such as clustering pieces of source data based on similarity) for the pieces of source data. In a number of embodiments, a plurality of pieces of source data are presented to the user and the user is asked to select similar pieces of source data, such as via grouping pieces of source data and/or other selection techniques as appropriate to the requirements of specific applications in accordance with embodiments of the invention. In many embodiments, the user is asked to select the two most similar pieces of source data within the presented source data (e.g. the user selects pairs of pieces of source data based on categorization criteria). In this way, by presenting different sets of source data the user annotations can be utilized to identify clusters of related pieces of source data based on the pairwise relationships identified in the user annotations. In a variety of embodiments, the annotations are performed using distributed data categorization server system **110**.

Distributed data categorization systems in accordance with embodiments of the invention are described above with respect to FIG. **1**; however, any of a variety of distributed data categorization systems can be utilized in accordance with embodiments of the invention. Systems and methods for distributed data categorization in accordance with embodiments of the invention are described below.

Distributed Data Categorization Server Systems

Distributed data categorization server systems are configured to assign pieces of source data from a set of source data to data categorization devices, receive annotations identifying clusters of source data within the set of source data from the data categorization devices, and determine categories for the pieces of source data based on the received annotations. A

distributed data categorization server system in accordance with an embodiment of the invention is conceptually illustrated in FIG. **2**. Distributed data categorization server system **200** includes processor **210** in communication with memory **230**. Distributed data categorization server system **200** also includes network interface **220** configured to send and receive data over a network connection. In a number of embodiments, network interface **220** is in communication with the processor **210** and/or memory **230**. In several embodiments, memory **230** is any form of storage configured to store a variety of data, including, but not limited to, data categorization application **232**, source data **234**, source data metadata **236**, and data categorization device metadata **238**. In many embodiments, source data **234**, source data metadata **236**, and/or data categorization device metadata **238** are stored using an external server system and received by distributed data categorization server system **200** using network interface **220**. External server systems in accordance with a variety of embodiments include, but are not limited to, database systems and other distributed storage services as appropriate to the requirements of specific applications in accordance with embodiments of the invention.

Data categorization application **232** configures processor **210** to perform a distributed data categorization process for set of source data **234**. The distributed data categorization process includes generating subsets of the set of source data **234** and transmitting the subsets of source data to one or more data categorization devices. In a variety of embodiments, the subsets of source data are transmitted via network interface **220**. In many embodiments, the selection of data categorization devices is based on data categorization device metadata **238**. As described below, the data categorization devices are configured to generate clusters of pieces of source data and generate source data metadata **236** containing annotations and/or other attributes for the pieces of source data based on the generated clusters. Source data attributes can include, but are not limited to, annotations provided for the piece of source data, the source of the provided annotations, and/or one or more categories identified as describing the piece of source data. In a variety of embodiments, data categorization application **232** configures processor **210** to perform the clustering and annotation processes. The distributed data categorization process further includes receiving the annotated pieces of source data and identifying categories containing one or more pieces of source data **234** based on the annotations and/or other attributes in source data metadata **236**. In a number of embodiments, the distributed data categorization process also includes identifying categories of source data that may contain additional sub-categories and iteratively performing the data categorization process on the identified categories to generate the sub-categories of source data.

In a number of embodiments, data categorization application **232** further configures processor **210** to generate and/or update data categorization device metadata **238** describing the characteristics of a data categorization device based on the pieces of source data provided to the data categorization device and/or the annotations generated by the data categorization device. Data categorization device metadata **238** can also be used to determine rewards and/or other compensation awarded to a data categorization device for providing annotations to one or more pieces of source data. Characteristics of a data categorization device include pieces of source data annotated by the data categorization device, the annotations applied to the pieces of source data, previous rewards granted to the data categorization device, the time spent annotating pieces of source data, demographic information, the location of the data categorization device, clustering criteria describ-

ing how the data categorization device clusters pieces of source data, and any other characteristic of the data categorization device as appropriate to the requirements of specific applications in accordance with embodiments of the invention.

Distributed data categorization server systems are described above with respect to FIG. 2; however, a variety of architectures, including those that store data or applications on disk or some other form of storage and are loaded into the memory at runtime can be utilized in accordance with embodiments of the invention. Processes for the distributed categorization of source data in accordance with embodiments of the invention are discussed further below. Distributed Data Categorization

The distributed categorization of source data involves the annotation of a set of source data by a number of data categorization devices. Using the annotations, categories describing the pieces of source data can be determined. These categories can be utilized to identify portions of the set of source data of interest for further analysis and/or category refinement. A process for distributed data categorization in accordance with an embodiment of the invention is illustrated in FIG. 3. The process 300 includes obtaining (310) a set of source data. Subsets of source data are determined (312) and assigned (314). Clusters within a subset of source data are generated (316). Categories of source data are identified (318). In many embodiments, if refinements to the categories are required (320), additional subsets of source data are determined (312).

In many embodiments, the obtained (310) source data contains one or more pieces of source data. The pieces of source data can be, but are not limited to, image data, audio data, video data, signal data, text data, or any other data appropriate to the requirements of specific applications in accordance with embodiments of the invention. The pieces of source data can include source data metadata describing attributes of the piece of source data. A variety of techniques can be utilized to determine (312) subsets of the source data as appropriate to the requirements of specific applications in accordance with embodiments of the invention, including determining overlapping subsets where every pair of pieces of source data occur in at least one subset, random sampling, and iterative sampling methods that adaptively choose maximally informative subsets. Additional processes for determining (312) subsets of source data that can be utilized in accordance with embodiments of the invention are described below. In a number of embodiments, determining (312) subsets of source data further includes generating instruction data describing how the source data should be clustered. The instruction data can be pre-determined and/or determined based on the attributes of the pieces of source data being provided. The determined (312) subsets are assigned (314) to one or more data categorization devices; a particular subset can be assigned to one or more data categorization devices. In a number of embodiments, a particular data categorization device can only provide one set of annotations for a particular subset of source data; other embodiments allow multiple annotations for a particular subset of source data by the same data categorization device. In several embodiments, the subsets are assigned (314) using data categorization device metadata describing the capabilities of the data categorization devices. In several embodiments, the capabilities of a data categorization are determined based on the received instruction data. In a variety of embodiments, a compensation value is associated with the assigned (314) subset; the compensation value can be pre-determined and/or determined dynamically based on the attributes of the source data, the instruction data, and/or the capabilities of the data categorization device.

Generating Subset Clusters

Subset clusters include one or more pieces of source data within an assigned (314) subset. Generating (316) clusters of source data includes associating pieces of source data based on the attributes of the source data, where pieces of source data in the same cluster have similar attributes. For example, pieces of source data can be clustered based on size, color, quantity, genre, location, or any other categorization criteria as appropriate to the requirements of specific applications in accordance with embodiments of the invention. The clusters can be generated (316) according to the instruction data and/or the capabilities of a data categorization device. In a variety of embodiments, the pieces of source data are annotated based on the differences between the generated (316) clusters; e.g. the data categorization device provides metadata describing the differences between the generated (316) clusters than can be utilized in identifying (318) categories based on the generated (316) clusters. In many embodiments, the generated (316) clusters are utilized to annotate the pieces of source data with binary pairwise labels for each pairing of pieces of source data within the assigned (314) subset such that each pair (a, b) of pieces of source data assigned (314) to a data categorization device j has a label l_{abj} where $l_{abj}=1$ if the pieces of source data are in the same cluster and $l_{abj}=0$ if the pieces of source data are in different clusters. In a number of embodiments, the annotations for the clusters of pieces of source data utilize pairwise distance and/or three-way comparisons between the pieces of source data. Other label values and labeling techniques can be utilized as appropriate to the requirements of specific applications in accordance with embodiments of the invention.

In several embodiments, the results of the generated (316) clusters are represented as a set of

$$\binom{M}{2}$$

binary labels (e.g. the annotations applied to the pieces of source data), where there are N total pieces of source data (indexed by i) and M pieces of source data in a determined (312) subset of source data; H subsets are determined (312) (indexed by h). In general, $M < N$, and in a number of embodiments, $M \ll N$. The subsets of source data are assigned (314) to J data categorization devices (indexed by j). The annotations received from all of the data categorization devices is the set of binary variables \mathcal{L} (indexed by $t \in \{1, \dots, T\}$), where

$$T = H \binom{M}{2}$$

is the total number of labels and l_t is the t-th label in \mathcal{L} . Associated with l_t is a quadruple (a, b, j, h_t) where $j_t \in \{1, \dots, J\}$ indicates the data categorization device producing the label, $a_t \in \{1, \dots, N\}$ and $b_t \in \{1, \dots, N\}$ indicate the two pieces of source data compared by the label, and $H_t \in \{1, \dots, H\}$ indicates the subset and data categorization device combination that generated the label.

Identifying Categories

Once clusters of pieces of source data have been generated (316) for the set of source data, categories can be identified (318). In a variety of embodiments, the generated (316) cluster information is received by a distributed data categorization server system from a plurality of data categorization

11

devices. Identifying (318) categories includes a number of indivisible groups that describe the generated (316) clusterings of pieces of source data across some or all of the subsets of source data. For example, suppose one data categorization device clusters objects into tall objects and another of short objects, while a second data categorization device clusters the same objects into a cluster of red objects and another of blue objects. Therefore, the identified (318) categories include tall red objects, short red objects, tall blue objects, and short blue objects.

In several embodiments, pieces of source data are represented as points in a Euclidian space and the data categorization devices are modeled as pairwise binary classifiers. Identifying (318) clusters is performed by clustering the source data points using a Dirichlet process mixture model, although a variety of discrete distributions can be utilized as appropriate to the requirements of specific applications in accordance with embodiments of the invention. Given a D dimensional vector x_i with components $[x_i]_d$ that encode source data piece i in the Euclidian space \mathbb{R}^D . A pairwise binary classifier is defined using the symmetric matrix $W_j \in \mathbb{R}^{D \times D}$ with entries $[W_j]_{d_1 d_2}$ and bias $\tau_j \in \mathbb{R}^D$. The likelihood of a data categorization device to generate (316) a particular clustering for images a_i and b_i with label l_i is

$$p(l_i | x_{a_i}, x_{b_i}, W_{j_i}, \tau_{j_i}) = \frac{1}{1 + \exp(-l_i A_i)}$$

where

$$A_i = x_{a_i}^T W_{j_i} x_{b_i} + \tau_{j_i}$$

Within the Euclidian space, pairs of vectors with common pairwise activity (or otherwise strongly correlated) over the points in the Euclidian space are identified (318) as being in the same category, while points that do not have common pairwise activity (or are otherwise not strongly connected) are identified (318) as being in different categories. The joint distribution of the points within the Euclidian space describing the pairwise activity can be given as

$$p(\Phi, V, Z, X, W, \tau, \mathcal{L}) =$$

$$\prod_{k=1}^{\infty} p(V_k | \alpha) p(\Phi_k | m_0, \beta_0, J_0, \eta_0) \prod_{i=1}^{\alpha} p(z_i | V) p(x_i | \Phi_{z_i}) \prod_{j=1}^J p(\text{vecp}\{W_j\} | \sigma_0^w) p(\tau_j | \sigma_0^{\tau}) \prod_{t=1}^T p(l_t | x_{a_t}, x_{b_t}, W_{j_t}, \tau_{j_t})$$

where the conditional distributions are defined as

$$p(V_k | \alpha) = \text{Beta}(V_k; 1, \alpha)$$

$$p(z_i = k | V) = V_k \prod_{i=1}^{k-1} (1 - V_i)$$

$$p(x_i | \Phi_{z_i}) = \text{Normal}\left(x_i; \mu_{z_i}, \sum_{z_i}\right)$$

$$p(x_i | \sigma_0^{\tau}) = \prod_d \text{Normal}([x_i]_d; 0, \sigma_0^{\tau})$$

12

-continued

$$p(\text{vecp}\{W_j\} | \sigma_0^w) = \prod_{d_1 \leq d_2} \text{Normal}([W_j]_{d_1 d_2}; 0, \sigma_0^w)$$

$$p(\tau_j | \sigma_0^{\tau}) = \text{Normal}(\tau_j; 0, \sigma_0^{\tau})$$

$$p(\Phi_k | m_0, \beta_0, J_0, \eta_0) = \text{Normal-Wishart}(\Phi_k; m_0, \beta_0, J_0, \eta_0)$$

and where

$$(\sigma_0^{\tau}, \sigma_0^w, \sigma_0^w, \alpha, m_0, \beta_0, J_0, \eta_0)$$

are fixed hyper-parameters.

In a number of embodiments, the joint distribution is inferred using a Variational Bayes method with a proxy distribution

$$q(\Phi, V, Z, X, W, \tau) = \prod_{k=K+1}^{\infty} p(V_k | \alpha) p(\Phi_k | m_0, \beta_0, J_0, \eta_0)$$

$$\prod_{k=1}^K q(V_k) q(\Phi_k) \prod_{i=1}^N q(z_i) q(x_i) \prod_{j=1}^J q(\text{vecp}\{W_j\}) q(\tau_j)$$

With parametric distributions

$$q(V_k) = \text{Beta}(V_k; \xi_{k,1}, \xi_{k,2})$$

$$q(\Phi_k) = \text{Normal-Wishart}(m_k, \beta_k, J_k, \eta_k)$$

$$q(x_i) = \prod_d \text{Normal}([x_i]_d; [\mu_i^x]_d, [\sigma_i^x]_d)$$

$$q(\tau_j) = \text{Normal}(\tau_j; \mu_j^{\tau}, \sigma_j^{\tau})$$

$$q(z_i = k) = q_{ik}$$

$$q(\text{vecp}\{W_j\}) = \prod_{d_1 \leq d_2} \text{Normal}\left([W_j]_{d_1 d_2}; [\mu_j^w]_{d_1 d_2}, [\sigma_j^w]_{d_1 d_2}\right)$$

and variational parameters

$$\{\xi_{k,1}, \xi_{k,2}\}$$

and

$$\{m_k, \beta_k, J_k, \eta_k\}$$

associated with the k-th mixture component. $q(z_i=k)=q_{ik}$ is the factorized assignment distribution for the i^{th} piece of source data with mean μ_i^x and variance σ_i^x associated with the point in Euclidian space. Similarly, μ_j^w and σ_j^w are the mean and variance for data categorization device j and μ_j^{τ} and σ_j^{τ} are the mean and variance for the data categorization devices' bias τ_j . The variational parameters can be determined utilizing a variety of functions as appropriate to the requirements of specific applications in accordance with embodiments of the invention, including Jensen's inequality. Using Jensen's inequality,

$$\log p(\mathcal{L} | \sigma_0^{\tau}, \sigma_0^w, \sigma_0^w, \alpha, m_0, \beta_0, J_0, \eta_0) \geq E_q \log p(\Phi, V, Z, X, W, \tau, \mathcal{L}) + \mathcal{H} \{q(\Phi, V, Z, X, W, \tau)\}$$

with entropy

$$\mathcal{H} \{q(\Phi, V, Z, X, W, \tau)\}$$

and Free Energy

$$E_q \log p(\Phi, V, Z, X, W, \tau, \mathcal{L}) + \mathcal{H} \{q(\Phi, V, Z, X, W, \tau)\}$$

13

The Free Energy can be approximated using a variety of distributions, including the unnormalized Gaussian function

$$g(\Delta_i) \exp\left\{\frac{(l_i A_i - \Delta_i)^2}{2\lambda(\Delta_i)(A_i^2 - \Delta_i^2)}\right\} \approx p(l_i | x_{a_i}, x_{b_i}, W_{j_i}, \tau_{j_i})$$

where

$$g(x) = (1 + e^{-x})^{-1} \text{ and } \lambda(\Delta) = [1/2 - g(\Delta)] / (2\Delta).$$

This results in the utility function

$$\mathcal{F} = E_q \log p(\Phi, V, Z, X, W, \tau) + \mathcal{H}\{q(\Phi, V, Z, X, W, \tau)\} +$$

$$\sum_i \log g(\Delta_i) + \frac{l_i}{2} E_q\{A_i\} - \frac{\Delta_i}{2} + \lambda(\Delta_i)(E_q\{A_i^2\} - \Delta_i^2)$$

describing the number of clusters of points in Euclidian space and, thereby, the identified (318) categories of source data based on the generated (316) subset clusters. Although a specific set of distributions and utility functions have been described above, a variety of distributions and utility functions, such as Gaussian mixture models, exponential distributions, Monte Carlo techniques, variational inference, and/or other techniques that enforce a scale for a latent space such that the points in space can be regularized and/or categorized (e.g. clustered) can be utilized as appropriate to the requirements of specific applications in accordance with embodiments of the invention.

In several embodiments, identifying (318) categories within the generated (316) subset clusters by identifying the boundaries between subset clusters. Identifying the boundaries between subset clusters includes observing the centroids of the generated (316) subset clusters and then selecting samples in the Euclidian space along the lines intersecting the centroids of the subset clusters and determining sets of points along (or near to) the line. Based on the similarity and/or differences of the identified points, the edges and boundaries of the subset clusters can be identified; the categories (318) are then identified based on the boundaries of the subset clusters. In a number of embodiments, the boundary determination is performed iteratively and refined as additional points are added to the Euclidian space and additional subset clusters and boundaries of subset clusters are identified. In many embodiments, metadata is associated with the subset clusters (and/or pieces of source data within the subset clusters) indicating the properties used to place the particular pieces of source data within a particular subset cluster is used to identify (318) the categories. This metadata can be provided by a data categorization device and/or be determined based on the similarities and/or differences between pieces of source data within a subset cluster or between multiple subset clusters.

Category Refinement

In a number of embodiments, once the categories have been identified (318), finer categorical distinctions exist in the pieces of source data associated with an identified (318) category. When it is desirable to identify these finer categorical distinctions, the categories can be refined (320). In many embodiments, categories are refined (320) based on attributes of the pieces of source data that are not described by the identified (318) categories. In a variety of embodiments, categories are refined (320) based on the number of pieces of data associated with an identified (318) category. In several embodiments, one or more of the clusters are clearly defined in Euclidean space, resulting in easily identified (318) categories. In other clusters, in particular larger clusters, it is

14

likely that there are several clusters present. These clusters within the larger cluster indicate the presence of sub-categories that can be refined (320) based on the initially identified (318) cluster. In a variety of embodiments, metrics related to the distribution of the samples around the centroid of a cluster are utilized to determine whether to attempt to further refine (320) a cluster to identify the sub-categories within the original cluster.

Although specific processes for the distributed categorization of source data are discussed above with respect to FIG. 3, any of a variety of processes, including those utilizing distributions differing from those described above, can be performed in accordance with embodiments of the invention. Processes for determining subsets of source data for distribution to and the characterization of data categorization devices in accordance with embodiments of the invention are discussed below.

Determining Subsets of Source Data

For a large set of source data, it may be impractical to analyze the entire set of source data at once in order to determine categories describing the pieces of source data in the set. By analyzing subsets of the set of source data, categories for the pieces of source data can be determined based on categories determined for the subsets. However, the subsets of source data need to be constructed carefully to ensure that sufficient overlap exists between subsets so that the subsets can be combined to determine categories for the whole while ensuring that every piece of source data is categorized. A process for determining subsets of source data for a set of source data in accordance with an embodiment of the invention is conceptually illustrated in FIG. 4. The process 400 includes obtaining (410) a set of source data. A subset size is determined (412) and source data is allocated (414) to the subsets. The subsets are completed (416).

In many embodiments, the set of source data is obtained (410) utilizing processes similar to those described above. The subset size is determined (412) such that a subset is smaller than the obtained (410) set. In a variety of embodiments, the subset size is determined (412) based on the capabilities of a data categorization device. The obtained (410) source data is allocated (414) deterministically to the subsets in a manner such that each piece of source data occurs in at least one subset and the subsets have an overlap in the pieces of source data allocated across all the subsets. In cases where the determined (412) subset size exceeds the allocated (414) pieces of source data, a subset is completed (416) by allocating additional pieces of source data within the set of source data that have not already been allocated (414) to the subset. In several embodiments, the completion (416) of a subset includes sampling without replacement from the set of source data.

By way of example, take set of N pieces of source data distributed into subsets having M pieces of source data in each subset. The N pieces of source data are allocated (414) across the subsets such that each subset contains

$$\left\lceil \frac{M}{V} \right\rceil$$

pieces of source data. The subsets are completed (416) by filling in the remaining

15

$$M - \left\lceil \frac{M}{V} \right\rceil$$

pieces of source data for each subset by sampling without replacement from the

$$N - \left\lceil \frac{M}{V} \right\rceil$$

pieces of source data that have not been allocated (414) to the subset. If R data categorization devices annotate each subset of source data, the total number of categorization tasks that need to be completed by data categorization devices is

$$H = R \left\lceil \frac{NV}{M} \right\rceil.$$

This results in

$$T = R \left\lceil \frac{NV}{M} \right\rceil \binom{M}{2} \in O(RNV M)$$

binary labels being applied as annotations to the pieces of source data.

Although a specific process for determining subsets of a set of source data is described above with respect to FIG. 4, a variety of processes, including those that utilize alternative sampling schemes, can be performed in accordance with embodiments of the invention. Processes for modeling the characteristics of data categorization devices in accordance with embodiments of the invention are discussed below.

Modeling Data Categorization Devices

A variety of attributes of a piece of source data can be utilized to cluster and categorize the piece of source data. Different data categorization devices employ different categorization criteria in the clustering of source data assigned to the data categorization device. These categorization criteria can lead data categorization devices to perform better or worse than other data categorization devices based on the attributes of the source data being categorized. By modeling the characteristics of the data categorization devices, source data can be targeted towards particular data categorization devices in order to improve the overall performance of the data categorization process. A process for annotating pieces of source data in accordance with an embodiment of the invention is conceptually illustrated in FIG. 5. The process 500 includes obtaining (510) a subset of source data. The source data is clustered (512) and annotated (514). In a variety of embodiments, characteristics of a data categorization device are modeled (516).

In a variety of embodiments, the subset of source data is obtained (510), clustered (512), and/or annotated (514) utilizing processes similar to those described above. In several embodiments, modeling (516) the characteristics of a data categorization device j include determining a predicted confusion matrix C_j where

$$[C_j]_{k_1 k_2} = E_q \left\{ \left[p(l=1 | x_a, x_b, W_j, \tau_j) p(x_a | \Phi_{k_1}) p(x_b | \Phi_{k_2}) \right] dx_a \right\}$$

and the predicted confusion matrix expresses the probability that the data categorization device j will assign source data k_1

16

and k_2 in the obtained (510) subset of source data to the same cluster (512). For the variational distributions $\Phi(k_1)$ and $\Phi(k_2)$, the expected values

$$E\{\Phi_{k_1}\} = \{m_{k_1}, J_{k_1}/n_{k_1}\}$$

and

$$E\{\Phi_{k_2}\} = \{m_{k_2}, J_{k_2}/n_{k_2}\}$$

results in the approximate confusion matrix

$$[\hat{C}_j]_{k_1 k_2} = g(\hat{\Delta}_{k_1 k_2}) \exp \{ (m_{k_1}^T u_j^w m_{k_2} + u_j^v - \hat{\Delta}_{k_1 k_2}) / 2 \}$$

that can be utilized in a variety of embodiments of the invention.

A specific process for annotating source data and modeling the characteristics for a data categorization device is described above with respect to FIG. 5; however, a variety of processes, including those that utilize alternative clustering techniques, can be utilized in accordance with embodiments of the invention.

Although the present invention has been described in certain specific aspects, many additional modifications and variations would be apparent to those skilled in the art. It is therefore to be understood that the present invention can be practiced otherwise than specifically described without departing from the scope and spirit of the present invention. Thus, embodiments of the present invention should be considered in all respects as illustrative and not restrictive. Accordingly, the scope of the invention should be determined not by the embodiments illustrated, but by the appended claims and their equivalents.

What is claimed is:

1. A method for labeling a set of source data, comprising:
 - obtaining a set of source data using a distributed data categorization server system comprising a processor and a memory connected to the processor;
 - determining a plurality of subsets of the source data using the distributed data categorization server system, where a subset of the source data comprises a plurality of pieces of source data in the set of source data;
 - providing the subsets of source data to at least one data categorization device, wherein:
 - the data categorization device comprises a processor, a memory, and annotator characteristic data describing the data categorization device; and
 - the data categorization device:
 - obtains a subset of source data;
 - generates a set of pairwise annotations for the obtained subset of source data, where a pairwise annotation indicates when a first piece of source data in the obtained source data is similar to a second piece of source data in the obtained source data based on the annotator characteristics; and
 - provides the pairwise annotations to the data categorization server system;
 - obtaining sets of pairwise annotations for each subset of source data using the data categorization server system;
 - identifying a category for each subset of source data based on the obtained pairwise annotations for the subset of source data and the annotator characteristics of the data categorization device assigned to the subset of source data using the distributed data categorization server system;
 - locating pieces of source data located in at least two of the subsets of source data using the data categorization server system;

17

generating source data metadata describing attributes for at least one of the located pieces of source data based on the categories assigned to each of the subsets of source data in which the located pieces of content are contained using the data categorization server system, where the source data metadata for a piece of source data describes attributes of the piece of source data; and

generating a taxonomy based on the identified categories and the set of source data using the distributed data categorization server system, where the taxonomy comprises relationships between the identified categories and the pieces of source data in the set of source data.

2. The method of claim 1, wherein a category in the taxonomy comprises one or more attributes of the pieces of source data associated with the category in the taxonomy.

3. The method of claim 1, further comprising iteratively identifying sub-categories for at least one identified category based on the pieces of source data associated with the identified category using the distributed data categorization server system.

4. The method of claim 3, wherein:

the at least one identified category is selected based on the attributes of the pieces of source data associated with the identified category; and

the identified sub-categories comprise at least one attribute from a piece of source data associated with the sub-category that is not present in the identified category.

5. The method of claim 1, further comprising generating instruction data using the distributed data categorization server system, where the instruction data describes the attributes of the pieces of the source data that should be used in generating the set of pairwise annotations.

6. The method of claim 5, wherein the instruction data is generated based on the attributes of the pieces of source data in the set of source data.

7. The method of claim 1, wherein the annotator characteristic metadata describes anticipated annotations based on the pieces of source data in the obtained subset of source data.

8. The method of claim 1, wherein clustering the set of source data into related subsets of source data further comprises:

generating a model comprising a set of points representing the pieces of source data in a Euclidian space using the distributed data categorization server system; and

clustering the set of points within the Euclidian space based on the set of pairwise annotations using the distributed data categorization server system.

9. The method of claim 8, further comprising estimating the number of clusters within the Euclidian space using the distributed data categorization server system.

10. The method of claim 1, wherein determining a plurality of subsets further comprises:

determining a subset size using the distributed data categorization server system, where the subset size is a measure of the number of pieces of source data assigned to a subset; and

deterministically allocating the pieces of source data to the determined subsets using the distributed data categorization server system.

11. The method of claim 10, further comprising allocating additional pieces of source data to the subsets using the distributed data categorization server system, where the additional pieces of source data are sampled without replacement from the set of source data not already assigned to the subset.

18

12. A distributed data categorization server system, comprising:

a processor; and

a memory connected to the processor and storing a data categorization application;

wherein the data categorization application directs the processor to:

obtain a set of source data;

determine a plurality of subsets of the source data, where a subset of the source data comprises a plurality of pieces of source data in the set of source data;

provide the subsets of source data to at least one data categorization device comprising a processor and a memory connected to the processor, wherein the data categorization device:

obtains a subset of source data;

generates a set of pairwise annotations for the obtained source data, where a pairwise annotation indicates when a first piece of source data in the obtained source data in the subset of source data is similar to a second piece of source data in the obtained source data based on the annotator characteristics; and

provides the pairwise annotations to the data categorization server system;

obtain a set of pairwise annotation for each subset of source data;

identify a category for each subset of source data based on the obtained pairwise annotations for the subset of source data and annotator characteristics of the data categorization device assigned to the subset of source data;

locate pieces of content contained in at least two of the subsets of source data;

generate source data metadata describing attributes for at least one of the located pieces of source data based on the categories assigned to each of the subsets of source data in which the located pieces of content are contained using the data categorization server system, where the source data metadata for a piece of source data describes attributes of the piece of source data; and

generate a taxonomy based on the identified categories and the set of source data, where the taxonomy comprises relationships between the identified categories and the pieces of source data in the set of source data.

13. The system of claim 12, wherein a category in the taxonomy comprises one or more attributes of the pieces of source data associated with the category in the taxonomy.

14. The system of claim 12, wherein the data categorization application further directs the processor to iteratively identify sub-categories for at least one identified category based on the pieces of source data associated with the identified category.

15. The system of claim 14, wherein:

the at least one identified category is selected based on the attributes of the pieces of source data associated with the identified category; and

the identified sub-categories comprise at least one attribute from a piece of source data associated with the sub-category that is not present in the identified category.

16. The system of claim 12, wherein the data categorization application further directs the processor to generate instruction data, where the instruction data describes the attributes of the pieces of the source data that should be used in generating the set of pairwise annotations.

17. The system of claim 16, wherein the instruction data is generated based on the attributes of the pieces of source data in the set of source data.

18. The system of claim 12, wherein the annotator characteristic metadata describes anticipated annotations based on the pieces of source data in the obtained subset of source data. 5

19. The system of claim 12, wherein the data categorization application further directs the processor to cluster the set of source data into related subsets of source data by:

generating a model comprising a set of points representing the pieces of source data in a Euclidian space; and clustering the set of points within the Euclidian space based on the set of pairwise annotations. 10

20. The system of claim 19, wherein the data categorization application further directs the processor to estimate the number of clusters within the Euclidian space. 15

21. The system of claim 12, wherein the data categorization application further directs the processor to determine a plurality of subsets by:

determining a subset size, where the subset size is a measure of the number of pieces of source data assigned to a subset; and deterministically allocating the pieces of source data to the determined subsets. 20

22. The system of claim 21, wherein the data categorization application further directs the processor to allocate additional pieces of source data to the subsets, where the additional pieces of source data are sampled without replacement from the set of source data not already assigned to the subset. 25

* * * * *

30